Eur Spine J (2005) 14: 1014–1026
DOI 10.1007/s00586-005-0911-9

ORIGINAL ARTICLE

Anne F. Mannion
Achim Elfering
Ralph Staerkle
Astrid Junge
Dieter Grob
Norbert K. Semmer
Nicola Jacobshagen
Jiri Dvorak
Norbert Boos

Outcome assessment in low back pain: how low can you go?

Received: 12 October 2004
Revised: 21 January 2005
Accepted: 11 February 2005
Published online: 4 June 2005

A. F. Mannion · A. Junge
D. Grob · J. Dvorak
Spine Unit, Schulthess Klinik,
Lengghalde 2, 8008 Zürich,
Switzerland

A. F. Mannion (✉)
Department Rheumatology
and Institute of Physical Medicine,
University Hospital,
Zürich, Switzerland
E-mail: anne.mannion@kws.ch
Tel.: +41-1-3857584
Fax: +41-1-3857590

A. Elfering · N. K. Semmer
N. Jacobshagen
Department of Psychology,
University of Berne,
Berne, Switzerland

R. Staerkle · N. Boos
Centre for Spinal Surgery,
University of Zürich,
Zürich, Switzerland

Abstract The present study examined the psychometric characteristics of a "core-set" of six individual questions (on pain, function, symptom-specific well-being, work disability, social disability and satisfaction) for use in low back pain (LBP) outcome assessment. A questionnaire booklet was administered to 277 German-speaking LBP patients with a range of common diagnoses, before and 6 months after surgical ($N = 187$) or conservative ($N = 90$) treatment. The core-set items were embedded in the booklet alongside validated 'reference' questionnaires: Likert scales for back/leg pain; Roland and Morris disability scale; WHO Quality of Life scale; Psychological General Well-Being Index. A further 45 patients with chronic LBP completed the booklet twice in 1–2 weeks. The minimal reliability (similar to Cronbach's alpha) for each core item was 0.42–0.78, increasing to 0.84 for a composite index score comprising all items plus an additional question on general well-being ('quality of life'). Floor or ceiling effects of 20–50% were observed for some items before surgery (function, symptom-specific well-being) and some items after it (disability, function). The intraclass correlation coefficient (ICC) ("test–retest reliability") was moderate to excellent (ICC, 0.67–0.95) for the individual core items and excellent (ICC, 0.91) for the composite index score. With the exception of "symptom-specific well-being", the correlations between each core item and its corresponding reference questionnaire ("validity") were between 0.61 and 0.79. Both the composite index and the individual items differentiated ($P < 0.001$) between the severity of the back problem in surgical and conservative patients (validity). The composite index score had an effect size (sensitivity to change) of 0.95, which was larger than most of the reference questionnaires (0.47–1.01); for individual core items, the effect sizes were 0.52–0.87. The core items provide a simple, practical, reliable, valid and sensitive assessment of outcome in LBP patients. We recommend the widespread and consistent use of the core-set items and their composite score index to promote standardisation of outcome measurements in clinical trials, multicentre studies, routine quality management and surgical registry systems.

Keywords Outcome measures · Psychometric properties · Low back pain · Spine surgery

Introduction

The notion that the outcome of treatment for musculoskeletal disorders should no longer be assessed solely (or even primarily) on the basis of imaging results, technical success, or objective functional/physiological measures is gaining increasing recognition. It is generally acknowledged that the success of any intervention for back pain should also be judged in relation to the patient's perception of the benefit gained—usually in terms of its effect on their pain, disability in everyday activities, work-capacity, quality of life, etc. [13]. Such measures are indispensable, as they assess the factors that will, ultimately, determine whether the patient is able to resume a normal working and social life again, or will, instead, continue to make use of the available healthcare resources.

The relative importance of the different dimensions of outcome to any given patient is difficult to estimate and is most likely dependent upon certain demographic characteristics of the patient (age, work-status, habitual activity level) and his or her main symptoms (pain, neurological deficit, deformity, functional disability) as well as the specific treatment administered. This means that, if a comprehensive evaluation of treatment efficacy is to be made, a range of different instruments must be used. However, when batteries of long questionnaires are administered, patient-compliance is likely to decline [17]. For large-scale quality management studies or studies of treatment effectiveness, it is often more important to examine the perceptions of the majority in regard to a few key issues than to examine the outcome of just a select few in great detail. This then requires the use of short, concise questionnaires, which nonetheless display the necessary psychometric characteristics of a valid outcome instrument.

To meet these needs, an international group of experts in primary care research put together a parsimonious six-item core-set of questions that would be practical for use in a wide variety of settings, including routine clinical care, quality management and more formal research [14]. The six-item "core-set" comprised several dimensions of outcome, each with a single item: pain severity (separately for back and leg); function; symptom-specific well-being; disability (work); disability (social role); plus 'satisfaction' with treatment. Each of the questions had been studied and validated elsewhere, sometimes as a component of other longer questionnaires [14]. The authors conceded that more data on the responsiveness of these measures and their application in languages other than English was required before they were implemented on a regular basis. However, there are still no reports in the peer-reviewed literature (other than in abstract form [30]) of the psychometric properties of the core-set per se. The present study sought to examine the test–retest reliability, validity, and responsiveness of the individual core-measures, and of a combined core measures composite index, in a large group of German-speaking low back pain (LBP) patients.

Methods

Study population

Patients were recruited from the spine centres of two neighbouring orthopaedic hospitals and from two chiropractic clinics. Inclusion criteria were low back and/or leg pain for a duration of at least 1 month and fluency in the German language. Exclusion criteria were severe medical problems (e.g. tumour, infection) or acute spinal injuries. During their visit to the hospital/chiropractic clinic for a consultation, all patients were invited to complete a questionnaire booklet. A total of 388 patients agreed to do so. Of these, 348 were sent a follow-up questionnaire booklet 6 months later (40 were not sent a second, due to administrative errors in the recall system, death of the patient, the patient didn't go on to receive the planned operation, etc.). Out of the 348 patients, 277 completed and returned the follow-up questionnaire giving a return rate of 80%. Details of the demographic characteristics of the patients are shown in Table 1.

Questionnaires

The questionnaire booklet enquired, amongst other things, about the following:

- Sociodemographic variables (education, family status, living conditions, work status)
- Pain intensity (Likert scales for: back-related problem today, at worst, at best [18])
- Pain-related disability in everyday activities (Roland and Morris disability questionnaire [31]; German version [19]
- Psychological general well-being [12] (German version [8])

Table 1 Study-sample characteristics

Total number	277
Sex (male/female)	126/151
Age mean ± SD (range)	55.9 ± 15.9 (18.8–86.9)
Treatment (op./cons.)	187/90
Diagnosis/number of patients	
Disc herniation	44
Spinal stenosis	100
Discopathy	39
Facet syndrom	12
Segmental instability	37
Non-specific low back pain	45

Table 2 Domain-index items (with German equivalent) and reference scales with number of items, response format and reference

Domain	Core-index item and response format	Reference scales and response format	References
Pain symptoms	Max of two-item values: "How severe was your back pain in the last week?" "How severe was your leg pain in the last week? *"Wie stark waren Ihre Rückenschmerzen in der letzten Woche?"* *"Wie stark waren Ihre Beinschmerzen in der letzten Woche?"* Response format, visual analogue scale: 0 (=no pain) to 10 (=worst pain I can imagine) *0 (=keine Schmerzen), 10 (=stärkste Schmerzen, die ich mir vorstellen kann)*	Three items from the McGill Pain Questionnaire Response format, Likert scale: 1 (=no pain) to 6 (=unbearable pain)	Melzack [28] Boos et al. [7]
Back function	"During the past week, how much did your back problem interfere with your normal work (including both work outside the home and housework)?" *"Wie stark haben Ihre Rückenbeschwerden Ihre normalen Aufgaben (Arbeit und zu Hause) in der letzten Woche beeinträchtigt?"* Response format, Likert scale: (1) not at all (2) a little bit (3) moderately (4) quite a bit (5) extremely *(1) gar nicht (2) ein wenig (3) mässig (4) erheblich (5) sehr stark*	(1) Roland Morris disability scale (24-item scale) Dichotomous response format: 0 (=does not apply); 1 (=applies) (2) Quality of life-physical subscale (=seven items from WHOQOL-Bref) Response format, Likert scale: one (=completely/an extreme amount) to five (=not at all) or one (=very satisfied) to five=(very dissatisfied)	Roland and Morris [31] Exner and Keel [19] WHOQOL GROUP [36, 37]
Symptom-specific well-being	"If you had to spend the rest of your life with the symptoms you have right now, how would you feel about it?" *"Wie würden Sie sich fühlen, wenn Sie den Rest Ihres Lebens mit Ihren derzeitigen Rückenbeschwerden leben müssten?"* * Response format, Likert scale: (1) very dissatisfied (2) somewhat dissatisfied (3) neither satisfied nor dissatisfied (4) satisfied (5) very satisfied *(1) sehr unzufrieden (2) unzufrieden (3) weder zufrieden noch unzufrieden (4) zufrieden (5) sehr zufrieden*	(1) Quality of life-physical subscale (=seven items from WHOQOL-Bref) (see above) (2) WHOQOL-Bref (26-item scale) (see above)	WHOQOL GROUP [36, 37] WHOQOL GROUP [36, 37]
General well-being	"How would you rate your quality of life?" *"Wie würden Sie Ihre Lebensqualität beurteilen?"* * Response format, Likert scale: (1) very bad (2) bad (3) average (4) good (5) very good *(1) sehr schlecht (2) schlecht (3) mittelmässig (4) gut (5) sehr gut*	(1) General life satisfaction scale (=four items from Psychological General Well-Being Index) Response format, Likert scale: from one to six (2) WHOQOL-Bref (26-item scale) (see above)	Depuy[12] Bullinger et al[8] WHOQOL GROUP[36, 37]
Disability	Mean of two-item scores: "During the past 4 weeks, how many days did you cut down on the things you usually do (work, housework, school, recreational activities) because of your back problem?" "During the past 4 weeks, how many days did your back problem keep you away from going to work (job, school, housework)?" *"An wievielen Tagen im letzten Monat haben Ihre Rückenbeschwerden (Kreuz- und Beinschmerzen) Sie gezwungen, Ihre gewohnten Tätigkeiten (Arbeit, Hausarbeit, Schule, Freizeitaktivitäten) einzuschränken?"* *"An wievielen Tagen im letzten Monat habe Ihre Rückenbeschwerden (Kreuz- und Beinschmerzen) Sie daran gehindert, zur Arbeit zu gehen (Arbeit, Schule, Hausarbeit)?"* Response format: Number of days/*Anzahl Tage* transformed into Likert scale where 1 (= 0 days) to 5 (= 22–31 days)	Roland Morris disability scale (24-item scale) (see above)	Roland and Morris [31] Exner and Keel [19]

Table 2 (Contd.)

Domain	Core-index item and response format	Reference scales and response format	References
Satisfaction with overall care	"Over the course of your treatment for your low back pain or leg pain (sciatica), how satisfied were you with your overall medical care?" *"Wie waren Sie bisher mit der Behandlung Ihrer Rückenschmerzen in unserem Spital zufrieden?"* Response format, Likert scale: (1) very dissatisfied, (2) somewhat dissatisfied, (3) neither satisfied nor dissatisfied, (4) somewhat satisfied, (5) very satisfied *(1) sehr unzufrieden, (2) unzufrieden, (3) weder zufrieden noch unzufrieden, (4) zufrieden (5) sehr zufrieden*		
Outcome of the treatment	"How would you rate the overall result of your back treatment/operation?" *"Wie beurteilen Sie das Ergebnis der Behandlung (Operation) Ihrer Rückenbeschwerden (Kreuz- und Beinschmerzen)?"* Response format, Likert scale: *(1) very good, (2) good, (3) satisfactory, (4) bad, (5) worse than before* *(1) sehr gut, (2) gut, (3) befriedigend, (4) schlecht, (5) schlechter als zuvor*	"To what extent has your back problem improved?" *"In welchem Umfang haben sich Ihre Rückenbeschwerden (Kreuz- und Beinschmerzen) verbessert?"* Response format, Likert scale: 0% (=no improvement) to 100% (=no complaints anymore) *keine Besserung, beschwerdenfrei*	
Improvement of symptoms			

Note. * scales inverted for each scale (such that 1="best result" and 5="worst result") before entered into the composite-index score

– Quality of life (WHOQOL-BREF: physical, psychological, social, environment) [36, 37]

The core-set questions of Deyo et al. [14] (see Table 2) were presented together with these reference questionnaires, at the beginning of the booklet.

At the 6-month follow-up, the above questions were presented again, along with three additional questions to assess: (1) satisfaction with overall medical care (as in the original set of Deyo et al [14]); (2) the global outcome of treatment (developed by ourselves to act

Table 3 Test–retest reliability results for each of the domain-index items and the full reference scales

Domain (core items, reference scale)	No items	Range*	M1	M2	t	*P*	ICC	95% CI_{ICC}	SEM	MDC95	MDC95%
Core Items											
(1) Pain	1	0–10	5.9	5.7	0.66	0.51	0.71	0.52–0.83	1.21	3.35	33.5
(2) Function	1	1–5	3.2	3.2	0.39	0.70	0.72	0.54–0.83	0.54	1.50	30.0
(3) Symptom-specific well-being	1	1–5	2.1	2.2	0.96	0.34	0.67	0.47–0.80	0.52	1.45	29.0
(4) General well-being	1	1–5	3.3	3.3	−0.27	0.79	0.80	0.65–0.88	0.37	1.02	20.4
(5) Disability	1	1–5	2.5	2.5	−0.17	0.86	0.95	0.91–0.97	0.33	0.90	18.1
(6) Core index	5	0–10	5.3	5.2	0.61	0.54	0.91	0.84–0.95	0.63	1.74	17.4
Reference scales (in brackets = the core-item(s) number that was cross-validated with the reference scale)											
Likert pain scales (1)	3	1–5	3.3	3.4	−0.51	0.61	0.85	0.75–0.92	0.34	0.94	18.8
Roland Morris disability (2) (5)	24	0–24	11.2	11.4	−0.26	0.80	0.78	0.63–0.88	2.50	6.91	28.8
WHOBREF physical sub-scale (2) (3)	7	1–5	3.2	3.3	−1.35	0.18	0.90	0.83–0.94	0.26	0.72	14.4
PGWB life satisfaction (4)	4	1–6	3.7	3.7	0.59	0.56	0.94	0.89–0.97	0.25	0.69	11.5
WHOBREF total score (3) (4)	26	1–5	3.7	3.7	−0.69	0.49	0.94	0.89–0.97	0.14	0.39	7.8

M1, *M2* mean value at baseline and at follow-up, respectively; *t* t-score from the *t*-test analysis; *P* significance of difference between mean values on the two occasions; *ICC* (intraclass correlation coefficient) (MS between participants−MS within participants)/((MS between participants + MS within participants)×(n−1)); CI_{ICC} 95% confidence intervals for the ICC; *SEM* (standard error of measurement) (SEM = SD baseline×SQR(1−rtt (where rtt = test–retest correlation, determined from the ICC)); *MDC95* minimum detectable change score; *MDC95%* MDC95 as percentage of maximum score.*see Table 2 for interpretation of score ranges

as an external criterion for treatment outcome, with which to assess the sensitivity to change of the core-set questions); (3) the percentage improvement in the back problem (also developed by ourselves, to be used to cross-validate the results of the category answers derived from the second global outcome question) (see Table 2).

All the full questionnaires to be used to assess the convergent validity of the core measures had been validated in the German language in previous studies. For the core-set questions, validated German versions for the pain visual analogue scales (VAS) and for function were already available from our previous questionnaire-validation studies [27, 32]; the remaining four items were translated by two bilingual native German speakers working together and then back-translated by a bilingual native English speaker to ensure that the translation accurately reflected the original English versions [22].

Examination of the psychometric properties of the questionnaires

Reliability assessment

The test–retest reliability was assessed in a group of 45 patients with chronic LBP (>3 months) who completed the questionnaire booklet twice over a period of 1–2 weeks (questionnaires sent out twice by mail).

Convergent validity of the core-set measures

Table 2 summarises the comparisons conducted with the reference questionnaires to examine the convergent validity of the core items.

The scores from the core item pain (maximum score from either back pain or leg pain 0–10 VAS) were compared with the average values from three 6-category Likert scales (for ''pain today'', ''worst pain in the last week'', ''least pain in the last week'').

The scores from the core item back function were compared with the Roland Morris disability scores and with the scores from the physical sub-scale of the WHOQOL-BREF questionnaire.

The scores from the core item symptom-specific well-being were compared with those from the physical sub-scale of the WHOQOL-BREF questionnaire. The scores from an additional item that we believed might improve the existing core-set, general well-being (a single question from the WHOQOL-BREF questionnaire, enquiring about overall quality of life), was compared with the scores of the general life satisfaction sub-scale of the psychological general well-being (PGWB) index. The scores for each of the well-being questions were also compared with the whole score of the WHOQOL-BREF questionnaire. For the disability item from the core-set (number of days 'out of action'), it was difficult to find a direct reference questionnaire. However, we examined the scores for this in relation to the Roland Morris disability score, on the assumption that the more disabled a patient is, the more absence from work he/she is likely to exhibit. The mean values from the ''disability (social role)'' and ''disability (work)'' were used, as many non-working patients had not completed the work question, and in those patients who completed both, the scores were often relatively similar.

From the five core-set items (pain, function, symptom-specific well-being, general well-being, disability) a composite index score was constructed. Firstly, all scales were linearly transformed into a 0–10 format. Pain intensity was already measured in this format, whilst function, and symptom-specific and general well-being were measured with a 1–5 point Likert scale (transformed according to the formula: category score marked by the patient -1×2.5). Disability (work) and disability (social role) were measured in days of work incapacity/restricted activity over the last month and could theoretically range from 0 to 31. These were firstly recoded into five categories to provide a similar scale as for the other items: (1) 0 days, (2) 1–7 days (3) 8–14 days (4) 15–21 days (5) >22 days, before being transformed to the 0−10 format.

The five transformed core-item scores were averaged to form a composite core index that ranged from 0 to 10.

Data analysis and statistics

Test-retest reliability

Paired *t*-tests were used to examine the significance of the difference in mean values for the two completions of all the questionnaires. The intraclass correlation coefficient (ICC) and the standard error of measurement (SEM) (or typical error of measurement) for the repeated trials, each with their 95% confidence intervals, were also determined. The SEM can be used to indicate the ''minimum detectable change'' ($MDC_{95\%}$) for the scale i.e. the degree of change required in an individual's score, in order to establish it (with a given level of confidence) as being a ''real change'', over and above measurement error [2] [24]. At the 95% confidence level, this is defined as $1.96\times\sqrt{2}\times SEM$ which is equivalent to $2.77\times SEM$.

The *internal consistency/internal reliability* for each full scale was assessed with Cronbach's alpha (for both baseline and follow-up questionnaires) and for the individual items, with estimates of the minimal internal reliability. Cronbach's alpha indicates the strength of the relationship between all the items within the test instru-

ment i.e. it examines the extent to which the instrument measures a single trait or characteristic. When a scale has only one item, there is no obvious way to calculate Cronbach's alpha. However, when there is a scale available that measures the same construct, the correlation with this scale can be used to calculate the minimal reliability of the single item [29] [34]. The resulting value represents the lower bound (i.e. a conservative estimate) of the internal reliability coefficient (minimal internal reliability) [29] [34].

Tests of *construct validity* indicate the extent to which an instrument's scores relate to those of other instruments in the way that one would expect, indicating that the instrument is really measuring the construct it is supposed to measure. This was assessed using two approaches. In the first approach, the relationship between each of the core-set items and its corresponding full scale was examined using Pearson product–Moment correlations. In the second approach, the significance of the differences between the mean scores of the conservative and surgical patients, for each of the core-set items and each of the full questionnaires, were examined using analysis of covariance (controlling for age and gender). This is based on the premise that surgical patients should have higher values for pain, disability, etc. than patients for whom conservative treatment was considered sufficient.

Floor and ceiling effects refer to the proportion of patients that obtain the lowest or highest possible score for the given scale, and for whom any transition to an even more extreme status would therefore not be measurable with that scale.

Responsiveness indicates the ability of an instrument to detect small but clinically important changes [15]. Responsiveness was calculated for the individual core-set questions, the whole core-set index and the full-scale questionnaires, using three different methods. Firstly, paired *t*-tests were used to examine the significance of the change in group mean scores from pre-treatment to 6 months post-treatment. Secondly, the *effect size* for the change in score was calculated by taking the mean of all the individual changes in score and dividing this by the standard deviation of these change scores [4]. Generally, an effect size of 0.2 is considered small, 0.5 moderate and 0.8 large [9]. Thirdly, the sensitivity and specificity of the given score relative to the patient global outcome was examined using the receiver operating characteristics (ROC) method [15]. Determining instrument responsiveness can be considered analogous to evaluating a diagnostic test, in which the instrument is the diagnostic test and the global outcome represents the gold standard [15]. The ROC curve synthesises information on sensitivity and specificity for detecting improvement according to some dichotomised, external criterion. It consists of a plot of 'true positive rate' (sensitivity) versus 'false positive rate' (1−specificity) for each of several possible cut-off points in change score [15]. Thus, sensitivity and specificity are calculated for a change score of 1 point, 2 points and so on. In the present study, the 5 global outcome categories for the overall result of the treatment were collapsed to provide a dichotomous outcome variable: "good outcome" (very good, good) and "poor outcome" (satisfactory, bad, worse than before). It was considered that, for elective procedures, satisfactory was not really a clinically good outcome, and hence the cut-off point for a good outcome was placed above this. The area under the ROC curve was interpreted as the probability of correctly discriminating between patients with a good and a poor outcome, using the change in the questionnaire scores; the area can range from 0.5 (no accuracy in discriminating) to 1.0 (perfect accuracy in discriminating). The ROC curve was also used to indicate the cut-off score change for distinguishing between good and poor outcomes [16]. This was determined using the simple approach of minimising errors (equivalent to maximising the sum of the specificity and sensitivity) [1], and quantified with the Youden index [38]. Statistical significance was accepted at the $P < 0.05$ level.

Results

Test–retest reliability

There was no significant difference between the mean scores on the two test occasions for any of the core items or for the full reference scales (Table 3).

Most of the individual core items showed good test–retest reliability: the intraclass correlation coefficients (ICC) ranged from 0.67 (symptom-specific QoL) to 0.95 (disability). The ICC for the 0–10 core index score was 0.91 (95% CI 0.83–0.95) and the standard error of measurement (SEM) was 0.63. The "minimum detectable change" ($MDC_{95\%}$) for the core index was thus calculated to be 1.7 points. For all the full reference scales, with the exception of 'disability', the test–retest reliability was slightly higher than for the corresponding single item (Table 3).

Internal consistency/minimal internal reliability

The estimate of minimal internal reliability for symptom-specific well-being was low, ranging from 0.07 to 0.26; this was the result of this item not correlating at all with its reference questionnaire (see below; Construct validity), as a good correlation between these two is required in order to estimate the minimal internal reliability for a one-item scale (see Methods). For all other core items, the estimates of minimal internal reliability ranged from 0.41 (function) to 0.78 (pain symptoms)

Table 4 Internal reliability of, and correlation between, the domain single items and the reference questionnaires (whole-group sample)

Core-index items	Estimate of internal reliability "Cronbach's α[a] "		Reference scales	Cronbach's α[a]		r[b]	
	Baseline	6 mo FU		Baseline	6 mo FU	Baseline	6 mo FU
Pain symptoms	0.78	0.69	Exner pain scale	0.74	0.91	0.76	0.79
Back function	0.51	0.61	*Roland and Morris*	0.88	0.92	0.67	0.75
	0.46	0.48	WHOQOL-BREF physical health	0.87	0.90	−0.63	−0.66
Symptom-specific well-being	0.11	0.26	WHOQOL-BREF physical health	(see above)		0.31	0.48
	0.07	0.19	WHOQOL-BREF whole score	0.89	0.94	0.25	0.42
General well-being	0.52	0.61	PGWB Gen. life satisfaction	0.82	0.88	0.65	0.73
	0.52	0.65	WHOQOL-BREF whole score	(see above)		0.68	0.78
Disability	0.42	0.51	*Roland and Morris*	(see above)		0.61	0.68
Index[c]	0.75	0.84					

[a]Estimate of Cronbach alpha reliability in single items calculated using the so called "attenuation formula" that estimates the true correlation between two scales that could be expected when measurements would include no measurement error (true rxy = rxy/(sqr(rxx×ryy)), where rxx and ryy are the reliability coefficients of both scales and rxy is the correlation coefficient between the index-item and the reference scale); the formula can also be applied when two measurements, one single item and a scale represent the same construct. In this case, expectation of the true correlation is 1. Setting the estimate of the true correlation (rxy) 1, one can estimate the Cronbach alpha of the item in calculating rxx = rxy2/ryy[29]

[b]Correlation between index-item and reference scale

[c]the index comprised five items: pain symptoms, back function, general well-being, symptom-specific well-being and disability

(Table 4). For the core item index score, the corresponding value was 0.75 in the baseline sample and 0.84 at the 6-month follow-up. For the reference scales, the Cronbach's alpha coefficients were all relatively high, ranging from 0.74 (Likert pain scale) to 0.94 (WHOQOL-BREF whole score).

Floor and ceiling effects

At baseline and at follow-up, some items showed notable floor and ceiling effects, although none to the extent cited as "adverse" (>70%) for health-related quality of life questionnaires [25]: 'function' demonstrated 31% ceiling effects and 20% floor effects at baseline and at follow-up, respectively; at follow-up, 'disability' showed 41% floor effects; and at baseline, symptom-specific well-being showed 50% floor effects (Table 5). All remaining items as well as the composite index demonstrated floor or ceiling effects less than 20% (Table 5).

For the two questions on satisfaction with treatment and global treatment outcome, there were in each case approximately 30% ceiling effects.

Construct validity

Correlation between the core items and the corresponding reference scales

The core items showed a moderate to high correlation with their corresponding reference questionnaires ($r = 0.60$–0.79; Table 4), with the exception of the item symptom-specific well-being: the correlations between the latter and the chosen reference scales (the WHOQoL physical quality of life, and WHOQoL whole score) were significant but very low ($r = 0.31$ and $r = 0.25$, respectively). Symptom-specific well-being also showed only a minimal correlation with the majority of other scales examined explicatively: with Roland Morris, $r = 0.25$; PGWB life satisfaction, $r = 0.20$; WHOQoL total score, $r = 0.25$. It exhibited the highest correlation with the Likert pain scale ($r = -0.39$ at baseline and -0.45 at follow-up).

Difference between surgical and conservative patients

Table 5 shows the mean scores for each of the core items and for the reference scales, for the surgical and conservative patients, at baseline and at 6 month follow-up. For all items and questionnaires, the surgical patients showed significantly more extreme values at baseline (i.e. worse pain, disability, well-being, etc.) than did the conservative patients, indicating that the instruments had good validity. Although the scores in both groups (surgical and conservative) generally improved after treatment, significant differences between the surgical and conservative patients still persisted for most of the domains at follow-up, with the exception of symptom-specific well-being. For many domains, the status of the surgical patients 6 months after their operation was similar to that of the conservative patients at baseline (i.e. before their treatment). There was a tendency (not significant) for a greater satisfaction with care and a higher rating of treatment outcome in the conservatively treated patients compared with the surgical patients.

Table 5 Mean scores (baseline and follow-up), responsiveness and floor/ceiling effects for the domain single items and the references questionnaires

Domain	Sample	Baseline mean (SD)	Follow-up mean (SD)	*P* baseline vs follow-up	Responsiveness[b]	Floor base	Ceiling baseline	Floor follow-up	Ceiling follow-up
Core-index items[a]									
Pain symptoms	All	6.7 (2.3)	4.3 (2.8)	< 0.001	0.87	1.9	7.1	8.7	2.6
	Surg	7.4 (1.8)	4.9 (2.8)	< 0.001	0.91	0.6	7.9	5.6	2.8
	Cons	5.5 (2.7)	3.0 (2.5)	< 0.001	0.81	4.5	5.6	14.9	2.3
P value, surg vs cons		< 0.001	< 0.001						
Back function	All	3.7 (1.2)	2.8 (1.3)	< 0.001	0.71	3.6	31.2	20.2	10.0
	Surg	4.1 (0.9)	3.1 (1.3)	< 0.001	0.68	0.5	35.5	13.3	13.8
	Cons	3.0 (1.3)	2.1 (1.0)	< 0.001	0.74	10.0	22.2	34.4	2.2
P value, surg vs cons		< 0.001	< 0.001						
Symptom-specific well-being	All	1.8 (0.9)	2.8 (1.2)	< 0.001	0.72	49.6	1.4	17.9	7.8
	Surg	1.7 (0.9)	2.7 (1.2)	< 0.001	0.80	53.8	1.6	18.5	5.6
	Cons	2.0 (1.0)	2.8 (1.3)	< 0.001	0.57	41.1	1.1	16.7	12.2
P value,surg vs cons		< 0.001	0.426						
General well-being	All	3.1 (1.1)	3.7 (0.9)	< 0.001	0.52	7.1	8.2	2.6	15.5
	Surg	2.9 (1.0)	3.5 (0.9)	< 0.001	0.59	9.4	6.1	2.2	12.6
	Cons	3.6 (0.9)	3.9 (0.8)	< 0.001	0.40	2.3	12.6	3.4	21.3
P value, surg vs cons		< 0.001	< 0.001						
Disability	All	3.0 (1.5)	2.1 (1.4)	< 0.001	0.60	18.5	29.8	41.1	13.7
	Surg	3.6 (1.4)	2.5 (1.5)	< 0.001	0.65	11.9	41.3	33.5	20.0
	Cons	2.1 (1.3)	1.5 (0.8)	< 0.001	0.51	30.7	9.1	54.7	2.3
P value, surg vs cons		< 0.001	< 0.001						
Index[c]	All	6.3 (2.0)	4.1 (2.4)	< 0.001	0.95	0.0	1.1	1.5	0.4
	Surg	7.1 (1.6)	4.7 (2.4)	< 0.001	0.98	0.0	1.6	1.7	0.0
	Cons	4.9 (2.0)	3.0 (1.7)	< 0.001	0.89	0.0	0.0	1.1	1.1
P value, surg vs cons		< 0.001	< 0.001						
Satisfaction with overall care	All	–	3.7 (1.3)		–	–	–	12.7	32.9
	Surg		3.6 (1.4)		–	–	–	14.3	30.3
	Cons		4.0 (1.2)		–	–	–	8.1	40.3
P value, surg vs cons			0.078						
Global outcome after treatment	All		2.3 (1.1)			–	–	30.6	2.4
	Surg		2.3 (1.2)			–	–	30.6	2.8
	Cons		2.1 (0.9)			–	–	30.6	1.4
P value, surg vs cons			0.162						
% Improvement in back problem	All		58.7 (29.9)			–	–	8.8	6.5
	Surg		55.8 (30.3)			–	–	11.2	5.6
	Cons		64.8 (28.4)			–	–	3.6	8.3
P value, surg vs cons			.037						
Reference scales									
Exner Pain Scale	All	3.7 (0.9)	2.7 (1.1)	< 0.001	1.01	0.4	1.1	9.3	0.4
	Surg	4.1 (0.8)	2.9 (1.1)	< 0.001	1.05	0.0	1.6	6.7	0.6
	Cons	3.1 (0.8)	2.2 (0.9)	< 0.001	0.92	1.1	0.0	0.0	4.4
P value, surg vs cons		< 0.001	< 0.001						
WHOQOL-BREF Phys Health	All	3.1 (0.8)	3.7 (0.8)	< 0.001	0.76	0.7	0.4	0.0	2.6
	Surg	2.8 (0.7)	3.5 (0.9)	< 0.001	0.89	1.1	0.0	0.0	4.4
	Cons	3.6 (0.8)	4.0 (0.7)	< 0.001	0.52	0.0	1.1	0.0	4.4
P value, surg vs cons		< 0.001	< 0.001						
WHOQOL-BREF Total well-being	All	3.7 (0.5)	3.9 (0.6)	< 0.001	0.50	0.0	0.0	0.0	0.0
	Surg	3.6 (0.5)	3.8 (0.6)	< 0.001	0.57	0.0	0.0	0.0	0.0
	Cons	4.0 (0.4)	4.1 (0.5)	< 0.001	0.34	0.0	0.0	0.0	0.0
P value, surg vs cons		< 0.001	< 0.001						
PDWB Life Satisfaction	All	3.6 (0.9)	4.0 (1.0)	< 0.001	0.47	0.4	0.0	0.0	0.7
	Surg	3.4 (0.9)	3.9 (1.0)	< 0.001	0.54	0.6	0.0	0.0	1.1
	Cons	4.0 (0.7)	4.3 (0.9)	0.002	0.36	0.0	0.0	0.0	0.0

Effect sizes

Table 5 shows the effect sizes (indicating the responsiveness or sensitivity to change of the measures) for each of the items and the reference questionnaires. For both conservative and surgical patients, the item general well-being had the smallest effect size (0.40 conservative, 0.59 surgical) and pain had the largest (0.81 conservative, 0.91 surgical).

Table 5 (Contd.)

Domain	Sample	Baseline mean (SD)	Follow-up mean (SD)	*P* baseline vs follow-up	Responsiveness[b]	Floor base	Ceiling baseline	Floor follow-up	Ceiling follow-up
P value, surg vs cons		< 0.001	0.005						
Roland and Morris disability questionnaire	All	13.5 (5.6)	9.4 (6.3)	< 0.001	0.78	0.4	1.1	4.1	0.4
	Surg	15.5 (4.6)	11.2 (6.0)	< 0.001	0.82	0.0	1.7	3.4	0.6
	Cons	9.5 (5.4)	5.7 (5.1)	< 0.001	0.69	1.1	0.0	5.6	0.0
P value, surg vs cons		< 0.001	< 0.001						

Satisfaction with treatment: "How satisfied were you with your overall medical care?"; Global outcome after treatment: "How would you rate the overall result of your back treatment/operation?"; Improvement: "To what extent has your back problem improved?"

[a]For scale ranges, see Table 2.

[b]Effect size for responsiveness was calculated by dividing the difference between the mean baseline and the mean follow-up score by the standard deviation of the difference in scores ((mean follow-up score−mean baseline score)/SD of the difference in scores; effect size convention labels are small (0.30), moderate (0.50) and large (0.80, cf. Cohen [9]).

[c]The index comprised five items: pain symptoms, back function, general well-being, symptom-specific well-being and disability. The five-item index was calculated as the mean of five ten-point (0–10) transformed item scores.

The effect size for the composite core index (all five items together) was large for both the conservative (0.89) and surgical (0.98) patients. The core index also had a larger effect size than any of the individual reference questionnaires, with the exception of the Likert pain scale (0.92 conservative, 1.05 surgical).

Receiver operating characteristics (ROC)

The proportion of patients in each global outcome category were as follows: *surgical* 30.6% very good, 27.2% good, 23.3% satisfactory, 16.1% bad, 2.8% worse than before; *conservative* 30.6% very good, 36.0% good, 30.6% satisfactory, 1.4% bad, 1.4% worse than before. The outcome determined using this Likert scale question showed a high correlation ($r = 0.73$) with the ratings on the percentage improvement scale, suggesting it had adequate construct validity. For the ROC analyses, the answers to the global outcome question were dichotomised as good (very good and good) and bad (satisfactory, bad, worse) (see Methods).

The area under the ROC curve was 0.77 (SE = 0.03, $P < 0.001$) for the whole group (Fig. 1), 0.67 for the conservative patients and 0.82 for the surgical group (each $P < 0.02$; Table 6). This indicates that the core index had good discriminative ability, especially for the surgical patients. The lower value for the conservative patients most likely arose because very few of them reported the result of treatment as "bad" or "worse than before" to enable reliable distinction between the different outcomes (i.e. it effectively became an analysis between "good" and "satisfactory").

Cut-off score using ROC analysis

The ROC curve was used to indicate the best cut-off in the change in core-index score for distinguishing between good and poor outcomes in the surgical patients only (Fig. 2). Optimising equally for both sensitivity and specificity, the cut-off point on the curve that is closest to the upper left corner of the ROC is selected, and the sensitivity and specificity at this point are given by the Youden index (J, J = sensitivity + specificity − 1).

A cut-off for the change in core-index score of 3 points (on the 0–10 scale) predicted a good outcome with a sensitivity of 65.7% and specificity of 94.4% (Youden index, 0.601). Although we maintain that "satisfactory" should be considered as a poor outcome, out of interest we repeated the analysis with "satisfactory" in the "good" outcome group. The cut-off score for discriminating between good and bad outcomes was in this case 1.85 (sensitivity of 62.8% and specificity of 93.9%; Youden index, 0.567), which was still slightly higher than the minimal detectable change value.

Discussion

General considerations

The assessment of outcome in patients being treated for low back pain has received much attention over the last two decades. An increasing number of domains have been identified as necessary for a comprehensive, multidimensional evaluation of outcome, and a wide array of instruments is now available for these purposes [6]. However, the corollary of this is that the number and length of the questionnaires to be completed by the patient becomes burdensome. In a drive to provide a practicable solution to the problem, Deyo et al. proposed the use of a parsimonious set of core measures for use in low back pain outcome assessment [14]. It was suggested that the consistent use of these core measures would improve standardisation, facilitate comparability among studies, allow pooling of data and promote the development of more multicentre studies. The idea was

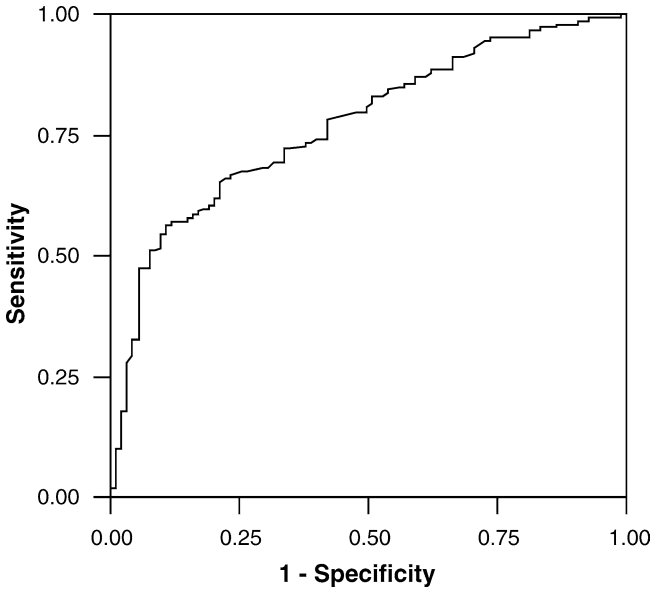

Fig. 1 Receiver operating characteristics curve for the core index for the whole group. See text for details

intuitively appealing, but no further studies on the psychometric properties of the core set were subsequently reported in the peer-reviewed literature (with the exception of one conference abstract [30]). This may explain the apparent hesitation of the scientific community to implement the core set in daily practice or research: despite widespread citation of the original paper (150 citations; Science Citation Index), to the best of our knowledge, there have been no subsequent research reports in the peer-reviewed literature in which the parsimonious core measures were employed as outcome measures. We have completed a thorough examination of the reliability, validity and responsiveness of the core measures in a large group of both conservative and surgical patients presenting with the most commonly encountered LBP-associated diagnoses.

The original six-item core-set covered several dimensions of outcome, including pain severity, function, symptom-specific well-being, disability (work, social role) and satisfaction with treatment. For measuring pain, the use of either a VAS or a Likert scale was originally recommended [14]; we chose to use the VAS within our core-set, as we anticipated that, in doing so, the data would yield greater comparability with many studies that have used this measure to date. Further, for the work disability questions, we elected to also include school, running the household, etc. as possible work options. We believed that this would reduce the number of missing answers in those patients who did not officially "work" in the sense of paid employment. This modification appears to have paid off, in that for fewer missing answers were observed in the present study (19.8 percent had at least one missing in any of the two disability items at baseline or at follow-up) in comparison with those reported by Pellise et al. [30] (52%), who presumably used the question as it was originally formulated. For those specifically interested in looking at the loss of paid working days, e.g. for the purposes of economic analyses, these individual disability items could simply be examined together with the patients' demographic data to extract those with a valid work status. In this way, no information is actually lost through our modification of the question. We chose to add one extra item to the pre-treatment core-set, namely an assessment of general well-being (quality of life). We felt that the latter was not adequately covered by the question on symptom specific well-being, yet quality of life is known to be an important attribute in musculoskeletal outcomes research [35]. And, indeed, the rather poor relationship between the questionnaire on symptom-specific well-being and the longer quality of life confirmed that these two are measuring somewhat different attributes. We also added a further question to the post-treatment set of items to directly assess the patient's impression of the success of the operation. We considered this to be quite a different concept from the question of satisfaction with treatment of the back problem in our hospital, which is not only influenced by the specific result of the treatment itself but also by the patient's perception of the level of care, the kindness and

Table 6 Area under the curve in receiver operating characteristic analyses

Sample	Area	Standard error[a]	Asymptotic Signal[b]	Asymptotic 95% confidence interval	
				Lower bound	Upper bound
Total	0.77	0.03	< 0.001	0.71	0.83
Surgery	0.82	0.03	< 0.001	0.75	0.88
Non-surgery	0.67	0.07	0.019	0.53	0.81

Test-result variable is changed in core-item index. Classification variable is good outcome of treatment
[a]Under the non-parametric assumption
[b]Null hypothesis: true area = 0.5

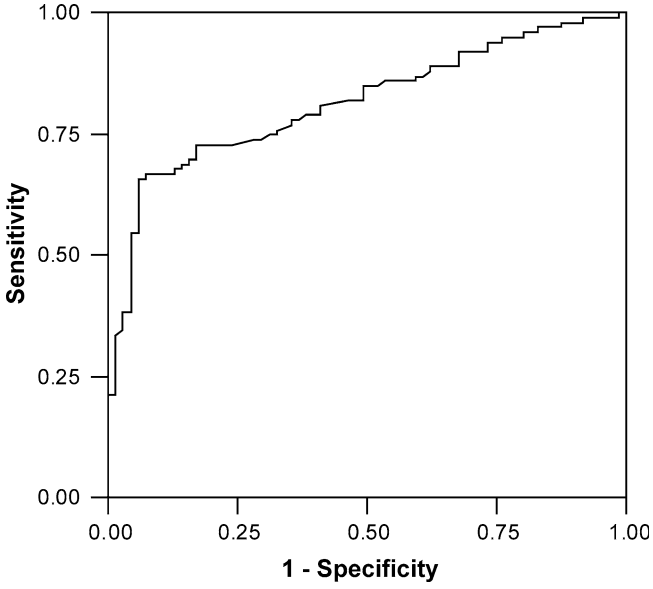

Fig. 2 Receiver operating characteristics curve for the core index for the surgical group only. See text for details

competence of the staff, the conditions in the hospital, etc. Again, this was borne out by the lack of any particularly strong relationship between the two satisfaction questions ($r = 0.12$, $P = 0.080$), suggesting that they were, indeed, delivering different information.

Although we chose to put all items together as a composite scale, in order to examine the feasibility of providing a single multidimensional measure, it is conceivable that the future user would prefer to examine change for each domain separately. This always remains an option, of course, although the corresponding psychometric properties for each item (e.g. for reliability, sensitivity to change, etc.) would have to be taken into consideration in interpreting individual change.

Our follow-up time was only 6 months, which by some standards is considered short, especially for surgical interventions. However, as the aim of the study was not to report on the outcome for specific procedures per se, but rather to examine the performance of the core measures in relation to corresponding but longer questionnaires, the time of follow-up was considered less relevant. Further, judging by the effect sizes recorded, 6 months was certainly long enough to detect clinically relevant changes in the patients' status after this time.

Test–retest reliability

Overall, the test–retest reliability of the core measures was good, with intraclass correlation coefficients (ICCs) for the individual items ranging from 0.67 to 0.95, and with an ICC for the whole core set (considered as one score) of 0.91. These values were well within the range previously reported for many longer outcome instruments [11, 20, 21]. The minimum detectable change (MDC_{95}) for the entire core index, determined from the test-retest analyses, was calculated to be 1.7 points (maximum possible score for the index = 10 points). This value represents the minimum difference in an individual's score required to state with 95% confidence that real change is responsible for the difference, as opposed to just measurement error (noise in the system). The value of 17% (i.e. 1.7 expressed as a percentage of the maximum score for the index) lay approximately in the middle of the range of the $MDC_{95\%}$ values for the full questionnaires (7.8–28.8%; Table 3) and is similar to that reported for other LBP outcome instruments [11]. The cut-off for predicting a good outcome was a change in core-index score of about 3 points (= 30% of the 0–10 scale). Hence, the clinically relevant change (30%; the "signal") for the core set far exceeded the minimum detectable change for the scale (17%; the "noise"), suggesting it may be superior to a number of other LBP outcome instruments in this respect [23]. Even if "satisfactory" was considered to be a good outcome, the clinically relevant change was 1.85 points, which still exceeded the $MDC_{95\%}$value.

Construct validity

With the exception of the item symptom specific well-being, the individual core items showed a moderate to high correlation with their corresponding full-version questionnaires ($r = 0.60$–0.79), indicating that they showed good concurrent validity. "Symptom-specific well-being" showed little relationship to any of these other measures. Whilst this made validation of this item difficult, we must conclude that it is perhaps delivering unique information that may be of importance to the multidimensional nature of the overall index. It was a relatively reliable (ICC 0.67) and sensitive (effect size 0.72) item, and so, for the time being, we recommend its continued inclusion in the core-set.

The validity of the core set was also shown by the fact that it discriminated significantly between the surgical and conservative patients at baseline. The overall condition of a surgical patient is expected to be considerably worse than that of a patient undergoing conservative treatment, and this was clearly detectable with the core set.

Floor and ceiling effects

For three of the individual core items (function, disability and symptom-specific well-being), there were notable floor and ceiling effects, although all remained

well below the "critical level" of 70% [25]. None of these effects were evident when the questions were combined to form the core index.

One factor that contributes to large floor and ceiling effects may be response bias. The selective participation of patients that were highly satisfied with their treatment may increase floor and ceiling effects. However, if great efforts are made to achieve good compliance at follow-up, this bias should be minimised. Since one way of improving compliance is to administer shorter questionnaires [17], the benefit of the core-set for capturing the response of a larger proportion of all patients becomes immediately apparent. Some authors maintain that, with Likert scales typical of the kind used in the core-set, scale sensitivity (resulting from floor and ceiling effects) may be a concern. It is argued that small but meaningful changes in the patient's condition may go unobserved when only five categories are presented. The use of scales with a higher resolution e.g. ten-point scales with end-point definitions (similar to the pain VAS) has thus been recommended [10]. In contrast, other studies of patients with musculoskeletal problems have shown that the Likert scales and 0–10 VAS yield almost identical results and are equally sensitive, and that the Likert scales have the added advantage of being easier to administer and interpret [5].

Responsiveness of the items/questionnaires

Good reliability and validity are prerequisites of all measurement instruments, especially when they are to be used to discriminate between subjects or predict prognosis [3, 26, 33]. However, the requirements for successful cross-sectional discrimination are not necessarily the same as those for successful longitudinal evaluation [26], and when measures are being considered for monitoring treatment outcome measures, it is essential to know how well they can detect small but important clinical changes, i.e. how responsive they are [16]. This information is essential not only for clinical decision making, but also for the determination of sample size in clinical trials, to ensure that the latter are adequately powered to detect a difference between treatments if one is present.

The effect sizes for the changes in the core index score after treatment were similar to the surgical and the conservative patients (0.98 and 0.89, respectively) and, according to most grading systems, would be considered as large [9]. With the exception of the Likert pain scale, the core index was the most responsive measure of all the questionnaires administered.

As perhaps expected, the symptom-specific items (e.g. pain, function, symptom-specific well-being) were generally more responsive than the more generic items (e.g. general well-being). Interestingly, disability due to back problems (days of cut-down activities and days off-work) was somewhat less responsive than the other back-specific items. This may have been because there were a number of individuals in the study who were not employed as such (e.g. homemakers), and who were thus, in the absence of any compensatory/cover system, still obliged to continue with their daily work activities. Had these patients been employed at a workplace where their absence would be covered by others and financially compensated, then changes in their work disability might have been more discernible.

In the surgical patients, a cut-off value for the reduction in the index score of approximately 3 out of ten points differentiated between a good and a bad global treatment outcome with a sensitivity of 66% and a specificity of 94%. The high specificity shows that few patients with an index change of >3 rated their outcome as poor. The somewhat lower value for the sensitivity indicates that some patients who indicated that they had improved according to the global outcome showed only a moderate improvement ($<3/10$ score reduction) in relation to their change in core-index score. This would suggest that there might be other influential factors, not currently included in the core-set, which contribute towards producing a "good" global outcome rating. We did not include the measure "satisfaction with treatment for the back problem" in the ROC analyses, as this item is only determined after surgery and therefore does not yield a change score. Arguably, general satisfaction with care—determined by the perceived degree of effort made by the clinician to alleviate the problem, or by the prior establishment of realistic expectations—could have a strong influence on governing the overall rating of treatment outcome. Perhaps an additional core question, concerning the patients' expectations before treatment and the extent to which these were fulfilled after treatment, would provide this missing link.

In conclusion, we have established that a slightly modified version of the core outcome measures for LBP, originally recommended by Deyo et al [14], displays psychometric characteristics that are to all intents and purposes as good as those of corresponding full-length questionnaires. Although some of the individual items show relatively large floor and ceiling effects, in contrast to the theory regarding the potential consequences of such effects these did not render the instrument unresponsive. We recommend the widespread and consistent use of the core-set in clinical trials, multicentre studies, routine quality management and surgical registry systems. Wider use of a uniform assessment tool would provide the framework for generating greater quantities of multinational LBP outcome data; ultimately, this should allow for an improved standard of care for the patient with LBP.

References

1. Altman DG, Bland JM (1994) Statistics notes: Diagnostic tests 3: receiver operating characteristic plots. BMJ 309:188
2. Beaton DE (2000) Understanding the relevance of measured change through studies of responsiveness. Spine 25:3192–3199
3. Beurskens AJ, de Vet HC, Köke AJ, van der Heijden GJ, Knipschild PG (1995) Measuring the functional status of patients with low back pain/assessment of the quality of four disease-specific questionnaires. Spine 20:1017–1028
4. Beurskens AJHM, de Vet HCW, Köke (1996) Responsiveness of functional status in low back pain: a comparison of different instruments. Pain 65:71–76
5. Bolognese JA, Schnitzer TJ, Ehrich EW (2003) Response relationship of VAS and Likert scales in osteoarthritis efficacy measurement. Osteoarthritis and Cartilage 11:499–507
6. Bombardier C (2000) Outcome assessments in the evaluation of treatment of spinal disorders: introduction. Spine 25:3097–3099
7. Boos N, Semmer N, Elfering A, Schade V, Gal I, Zanetti M, Kissling R, Buchegger N, Hodler J, Main CJ (2000) Natural history of individuals with asymptomatic disc abnormalities in magnetic resonance imaging. Predictors of low back pain-related medical consultation and work incapacity. Spine 25:1484–1492
8. Bullinger M, Heinisch M, Ludwig M, Geier S (1990) Skalen zur Erfassung des Wohlbefindens: Psychometrische Analysen zum "Profile of Mood States" (POMS) und zum "Psychological General Wellbeing Index" (PGWI). Zeitschrift für Differentielle und Diagnostische Psychologie 11:53–61
9. Cohen J (1988) Statistical power analysis for the behavioral sciences. Lawrence Earlbaum Associates, Hillsdale
10. Cummins RA, Gullone E (2000) Why we should not use 5-point Likert scales: The case for subjective quality of life measurement. Proceedings of the Second International Conference on Quality of Life in Cities National University of Singapore, Singapore:74–93
11. Davidson M, Keating JL (2002) A comparison of five low back disability questionnaires: reliability and responsiveness. Phys Ther 82:8–24
12. Depuy HJ (1984) The Psychological General Well-Being (PGWB) Index. In: Assessment of quality of life in clinical trials of cardiovascular therapies. Le Jacq, New York, pp 170–183
13. Deyo RA, Andersson G, Bombardier C, Cherkin DC, Keller RB, Lee CK, Liang MH, Lipscomb B, Shekelle P, Spratt KF, Weinstein JN (1994) Outcome measures for studying patients with low back pain. Spine 19:2032S–2036S
14. Deyo RA, Battie M, Beurskens AJHM, Bombardier C, Croft P, Koes B, Malmivaara A, Roland M, Von Korff M, Waddell G (1998) Outcome measures for low back pain research. A proposal for standardized use. Spine 23:2003–2013
15. Deyo RA, Centor RM (1986) Assessing the responsiveness of functional scales to clinical change: An analogy to diagnostic test performance. Journal of Chronic Diseases 39:897–906
16. Deyo RA, Diehr P, Patrick DL (1991) Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Controlled Clin Trials 12(Suppl):142–158
17. Edwards P, Roberts I, Clarke M, DiGuiseppi C, Pratap S, Wentz R, Kwan I (2002) Increasing response rates to postal questionnaires: systematic review. BMJ 324:1183–1191
18. Exner V (1998) Lebensqualität bei chronischen Rückenschmerzenpatientlnnen [Quality of life in chronic back-pain patients]. Unpublished Master's Thesis, University of Basel; Basel, Switzerland
19. Exner V, Keel P (2000) Erfassung der Behinderung bei Patienten mit chronischen Rückenschmerzen. Schmerz 14:392–400
20. Fritz JM, Irrgang JJ (2001) A comparison of a modified Oswestry Low Back Pain Disability Questionnaire and the Quebec Back Pain Disability Scale. Phys Ther 81:776–788
21. Gronblad M, Hupli M, Wennerstrand P, Jarvinen E, Lukinmaa A, Kouri JP, Karaharju EO (1993) Intercorrelation and test-retest reliability of the Pain Disability Index (PDI) and the Oswestry Disability Questionnaire (ODQ) and their correlation with pain intensity in low back pain patients. Clin J Pain 9:189–195
22. Guillemin F, Bombardier C, Beaton D (1993) Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. J Clin Epidemiol 46:1417–1432
23. Hagg O, Fritzell P, Nordwall A (2003) The clinical importance of changes in outcome scores after treatment for chronic low back pain. Eur Spine J 12:12–20
24. Hopkins WG (2000) Measures of reliability in sports medicine and science. Sports Med 30:1–15
25. Hyland ME (2003) A brief guide to the selection of quality of life instrument Health and Quality of Life Outcomes 1:24
26. Kirschner, Guyatt A (1985) A methodological framework for assessing health indices. J Chronic Dis 38:27–36
27. Mannion AF, Junge A, Fairbank JCT, Dvorak J, Grob D (2004) Development of a German version of the Oswestry Low Back Index. Part 1: cross-cultural adaptation, reliability, and validity. Eur Spine J. DOI: 10.1007/s00586-004-0815-0
28. Melzack R (1975) The McGill Pain Questionnaire: Major Properties and Scoring Methods. Pain 1:277–299
29. Nunnally JC (1978) Psychometric Theory, McGraw-Hill, New York
30. Pellise F, Alvarez L, Escudero O, Pont A, Ferrer M (2003) Metric characteristics of the six-question "core set" in the evaluation of back pain. Eur Spine J 12:S12–S13
31. Roland M, Morris R (1983) A study of the natural history of back pain. Part 1: Development of a reliable and sensitive measure of disability in low-back pain. Spine 8:141–144
32. Staerkle R, Mannion AF, Elfering A, Junge A, Semmer NK, Jacobshagen N, Grob D, Dvorak J, Boos N (2004) Longitudinal validation of the fear-avoidance beliefs questionnaire (FABQ) in a Swiss-German sample of low back pain patients. Eur Spine J 13:332–340
33. Stratford PW, Binkley J, Solomon P, Gill C, FInch E (1994) Assessing change over time in patients with low back pain. Phys Ther 74:528–533
34. Wanous JP, Hudy MJ (2001) Single-item reliability: a replication and extension. Organizational Research Methods 4:361–375
35. Ward MM (2004) Outcome measurement: health status and quality of life. Curr opin Rheumatol 16:96–101
36. WHOQOL group (1998) Development of the World Health Organization WHOQOL-BREF quality of life assessment. The WHOQOL group. Psychol Med 28:551–558
37. WHOQOL group (1998) The World Health Organisation WHOQOL-BREF quality of life assessment (WHOQOL): development and general psychometric properties. Soc Sci Med 46:1569–1585
38. Youdon W (1950) Index for Rating Diagnostic Tests. Cancer 3:32–35